W. Christopher Lenhardt
Research Scientist, Earth Data Science
Renaissance Computing Institute (RENCI)
The University of North Carolina at Chapel Hill
919-445-0480
clenhardt@renci.org

Thank you for the opportunity to provide our thoughts to the Committee on the provision of North Carolina data related to the National Pollutant Discharge Elimination System (NPDES), as well as access to complementary environmental data. Also, thank you to the North Carolina Policy Collaboratory for facilitating the engagement between RENCI and Department of Environmental Quality (DEQ) and Department of Information Technology (DIT).

As you may know RENCI has been tapped by the NC Collaboratory to lead the effort developing a proposal outlining potential avenues to facilitate access to NPDES data and information, avenues for digitization of non-digital permit material, as well as access to other relevant environmental data. I would like to use today as an opportunity to present an early look at what we have learned so far and to outline some of the key issues we expect to delve into more deeply as we prepare the report for Spring 2018.

Since the initial inquiry from the Collaboratory, we have held very productive initial meetings with representatives from DEQ and the relevant Department of Information Technology (DIT) staff responsible for permit-related information. The Collaboratory has facilitated these opportunities and these meetings have focused on developing an initial understanding of the types of data and information available, the current tools for accessing data, and plans for enhancements.

Our path forward is to work closely with DEQ to develop a greater understanding in the following areas. First, we would like to continue to develop our familiarity with pertinent DEQ data holdings, both born digital and legacy analog. As part of this high-level cataloging we will also need to understand the relationship between the various types of data and information. For example, when a permit is submitted, what are the other documents or data that are associated with processing the permit such as maps, field measurements, images, and so on. Second, we also plan to work with DEQ to develop a set of reasonable user scenarios related to accessing the data. For example, the committee staff provided the scenario of a “water treatment plant operator from City X want[ing] to find relevant data online”. Developing a limited set of these scenarios will help to frame the type of data and information needs which can also help to drive requirements for the types of information systems and information products to develop.

North Carolina collects significant amounts of data and information related to water permits and the environment. DEQ also has a number of systems already available to access these data.

At the same time, DEQ is also investing in geographic information systems, also known as GIS technology, to provide a map-based interface to view and retrieve permit, and permit-related data. As you may know, GIS, is an information technology that allows 'georeferenced' data to be linked to a map interface. Google Maps has popularized this approach to information visualization, management and retrieval.

Digitization

Another aspect we are working to address is to help explore pathways to address the not uncommon need to convert legacy records in analog form to digital. The digitization of analog records, of course, has its own set of challenges. As part of the preliminary work developing the proposal we are looking at elements of the digitization process that affect the usability of scanned materials and which can influence costs.

When converting analog documents to digital, the challenge isn't usually the actual conversion from analog to digital, except when the item is of unusual size or material. The challenge is the work needed to make the raw scanned images useful. For example, in addition to scanning a text document, the resulting digital image needs to be processed through an optical character recognition (OCR) process. OCR transforms a scanned image from an image format-- think .jpeg or .gif-- into a format where the individual characters and elements on the page are rendered as objects in the document. Conversion in this way, allows the content to be indexed through machine-based algorithms which reduces some of the manual work related to indexing.

Other requirements that are important aspects of digitization are quality control and quality assurance. When digitizing large volumes of documents, a system of quality assurance and quality control should be implemented to reasonably insure that what comes out of the digitization process is usable. The physical quality and physical format (e.g. loose single pages versus a bound volume) of the input documents will affect the ability to process and affect the quality of the output. The more demanding the input material, the more effort needed to render a usable product.

Finally, the usability of a digitized object is only as good as the system to index and add metadata to the object. Some of this can be automated and some can be done as part of the ingest process, i.e. developing and/or applying and indexing and cataloging system. Documents that can't be searched or identified are next to worthless from a usability standpoint.

Developing Access to Permit and Related Environmental Data

On the more general topic of providing access to data and information, the challenges can be synthesized along a few different dimensions.

The first is having a system or system of systems that can connect to or integrate with multiple data sources/systems using standard interoperability protocols and recognized data formats. DEQ DIT is already working on this through their efforts such as the web GIS platform.

A second dimension is providing a search capability usable by a non-specialist. The interface is not necessarily the problem, the challenge is having enough value-add on the data, e.g. metadata-- data about the data-- and documentation, to make it 'find-able', accessible, and usable. The value add work is also essential to developing catalogs of holdings that can be leveraged in multiple applications.

There is also a tension between general and specific. That is if you create a system to do one particular thing well, like enable a city water manager to find information pertinent to their particular need, it might make it harder to support other uses and vice-versa. The more you make a system that tries to support all kinds of uses the greater the risk of creating a system that is too complicated to maintain and use effectively. Finding the right balance in this context can be helped by developing a set of user stories and ideas for information products as mentioned earlier.

Unfortunately there is no single best answer. However, doing the work to standardize, add metadata, and catalog does go a long way to helping with these problems. Investments in this type of work, improves the ability to deliver a higher level of access and potentially provide more services on top of the data.

North Carolina Open Data

Finally, the question was raised asking where NC stands in terms of open access to digital data. As part of the proposal development we plan on doing more detailed research to develop a systematic understanding of how states in the US approach providing access to data. The state of North Carolina provides open access to a great array of data related to the state and state government, including DEQ's open data access portal. See http://data-ncdenr.opendata.arcgis.com/.

More generally, however, it does seem that many states have a portal devoted to providing a one-stop access to state data, for example https://data.texas.gov/ or https://data.virginia.gov/ or https://data.michigan.gov/.

The value of developing ways to make data generated under the auspices of the state of North Carolina should not be underestimated and when used these data can improve agriculture outcomes, support emergency response, facilitate effective management of valuable North Carolina natural resources, and promote health and well-being to name a few. For more on this aspect please see the recent report from the North Carolina Department of Commerce, Board of Science, Technology, and Innovation, NC in the Next Tech Tsunami: Navigating the Data Economy. (https://www.nccommerce.com/Portals/6/Documents/Resources/NC%20Big%20Data%20Report.pdf)

To conclude, we are looking forward to continuing to develop the relationship with DEQ regarding data access and to explore fruitful avenues for mutually beneficial collaboration. In the course of these meetings, we have already learned about data that DEQ has related to

Hurricane Matthew flooding that would be very useful to a watershed flood mapping project we successfully concluded this Spring funded by the Collaboratory. It should also be noted that the flood mapping project leveraged other high value data procured and held by North Carolina; the high-resolution LIDAR data, high resolution data that can be processed for terrain mapping.

We very much appreciate the opportunity provided by the Collaboratory to help North Carolina leverage the best available environmental science for practical use to benefit to North Carolina citizens.

We look forward to updating you in the near future as we continue to work on this project.

Thank you.